Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. 376, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds.

***Spitzer* Science Center within an Enterprise Architecture**

T. Handley and the IPAC ISG, DBA, and SDM Teams

Infrared Processing and Analysis Center and Spitzer Science Center, California Institute of Technology, Pasadena, CA, USA

Abstract. The *Spitzer* Science Center's (SSC) evolutionary development approach, coupled with a flexible, scaleable hardware and software architecture has been key in *Spitzer*'s ability to handle an explosion of data products, evolving data definitions, and changing data quality requirements. *Spitzer* is generating (depending on the campaign and instrument) about 10 TB of pre-archive data every 14 to 20 days. This generally reduces to between 3 TB and 6 TB of standard products, again depending on the campaign and instrument. This paper will discuss (1) the *Spitzer* Science Center's responses to evolving data, quality, and processing requirements and (2) how robust or not was the original architecture to allow *Spitzer* to accommodate on-going change.

1. *Spitzer* Mission - Background and Today

Detailed technical descriptions of *Spitzer* and its three focal plane instruments have been previously published (Werner et al. 2004, Roellig et al. 2004). In this paper, we will limit the background description to those features of *Spitzer* most directly related to driving the shape of the ground system.

The *Spitzer Space Telescope* is a multi-purpose observatory cooled passively and with liquid-helium for astronomical observations over six octaves of wavelength in the infrared. The liquid helium is now projected to last 5.2 years (Storrie-Lombardi 2005). *Spitzer* utilizes an Earth-trailing heliocentric orbit. As seen from Earth, *Spitzer* recedes at about 0.1 AU per year and will reach a distance of 0.62 AU in five years. For *Spitzer*, the Earth-trailing orbit has several major advantages over near-Earth orbits. The principal advantage is the distance from Earth and its heat; this facilitates the extensive use of radiative cooling, which makes *Spitzer's* cryo-thermal design extremely efficient. The orbit also permits excellent sky viewing and observing efficiency (Werner 2006).

There are three instruments MIPS (spectroscopy), IRS (spectroscopy), and IRAC (imaging). Each instrument consists of a cold assembly mounted in the cryostat and warm electronics mounted in the spacecraft bus. The science instruments cannot be used simultaneously; only one instrument can be powered on at a time. The CTA has an outer shell that radiates to cold space in the anti-Sun direction, and is shielded from the Sun by the solar panel assembly.

At launch plus three years, *Spitzer* has proven to be well behaved. As of the end of week 150 (2006 Oct 11 at 20:00:00 UT) *Spitzer* had executed 23,255 science astronomical observation requests (AORs) totaling 17,830 hours. Other important milestones include:

- 2004 May 11: Opened the Science Data Archive

- 2006 May: Completed the third general observer proposal call
- 2006 November: Issued cycle-4 call for proposals

Data may be discovered, selected and accessed via a downloadable archive interface, Leopard, or fetched from an anonymous ftp site. Recently, weekly download totals have been more than 1.4 TB. For further information, you may start at:

- `http://ssc.spitzer.caltech.edu/ost/`
- `http://ssc.spitzer.caltech.edu/tools/`
- `http://ssc.spitzer.caltech.edu/archanaly/`

Popular prepackaged products[1] are also available.

2. Process: SSC's Approach to Architecture, Design and Development

SSC chose an evolutionary approach (Kasse 2005) for its lifecycle process. This approach allowed SSC to deal with uncertainty and evolution in requirements and in technology—an especially important issue for the SSC with its long development phase (4–5 yr) and a long system life (8–11 yr). Level 3 requirements could be more abstract and therefore subject to interpretation at any point in time.

As Level 1 and 2 requirements were virtually immutable, Level 3 detailed requirements remained more abstract to give SSC some flexibility. Alternative solutions could be explored and could be pursued further as new technology options became available. An evolutionary approach meant that intermediate designs could be saved for future use. Thus, intermediate designs could be implemented as prototypes but never operationally implemented.

A functional architecture was used to partition and allocate requirements to different subsystems and operational modes. SSC set detailed requirements in the architecture, interfaces, and environment of the system - and allowed the detailed requirements of implementation to be more abstract. That meant SSC could proceed with a good understanding of what the system had to do functionally. This gave SSC designers enormous freedom. The choice of the evolutionary approach allowed new ideas to be validated early. It also allowed SSC to build a system that we knew could be scaled up by a factor of 10, even though that was *n*ot a requirement. It came in very handy, however, when the system was scaled up by a factor of more than 100.

The SSC systems engineer has said, "One of the biggest issues was that the few hard requirements we had were time-fixed." For instance, it was not sufficient to demonstrate a system that could be scaled to handle 23 TB of data. The complete, but empty, archive had to be in place and functional at launch—a fixed time. This was a Level 2 requirement. Had the requirement allowed for providing storage when it was needed, the SSC would have been far better positioned to take advantage of all-on-disk storage architectures.

For the development lifecycle, SSC chose a freeze, build, and deliver approach. When there is a need to build a system, the available solution that best

[1] http://data.spitzer.caltech.edu/popular/

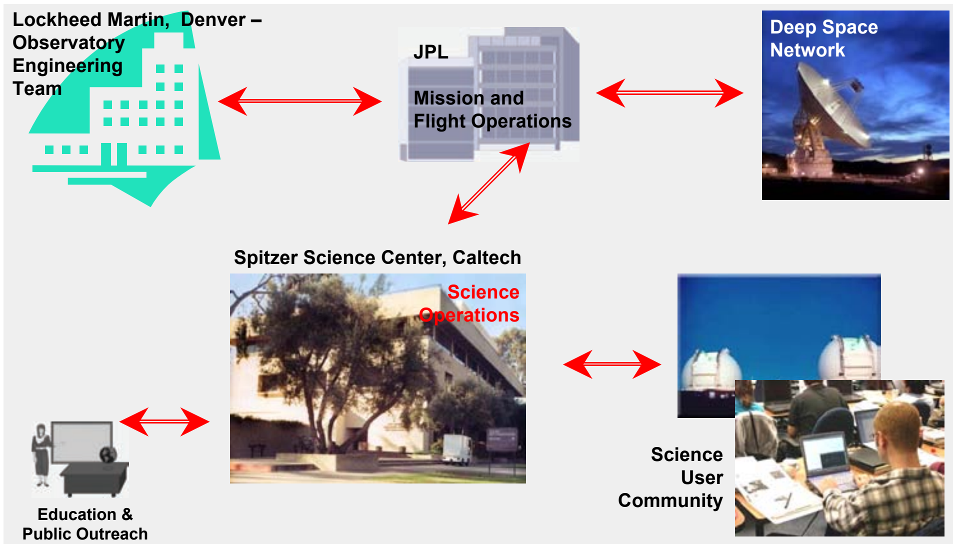

Figure 1. Where is the SSC? (Storrie-Lombardi 2005).

meets the current requirements is selected and implemented using the traditional engineering approach: incremental system deliveries. The first system release, dubbed S1.0, was to validate the data flows and interfaces between JPL Mission Operations and SSC Science Operations System (SOS). Starting around S4.0, the SSC settled into a regular pattern of 2 deliveries each year, nominally in May and November.

- S1.0 1998 December (SOS/FOS interfaces)
- S2.0 1999 October
- ...
- S15.0 2006 November 15.

3. SSC Science Operations System at Launch

Figure 1 illustrates that the SSC, its processes, its procedures, its teams, and its systems are part of an overall enterprise-level project information system. From Figure 2 one can see that at the enterprise level, the *Spitzer* Ground Data Systems (GDS) process flow is very traditional.

As shown in Figure 3, the SSC is responsible for evaluating and selecting observing proposals, providing technical and scientific support to the observer community, performing mission planning and science observation scheduling, science instrument calibration and science instrument performance monitoring, data processing and production of archival quality data products, and funding science research. The uplink process starts at the SSC with the generation of one week long observatory schedules, which are then sent to JPL for final command generation, validation and radiation to the observatory. Data are received from the observatory by the Deep Space Network (DSN) and transferred to JPL where level 1 data processing is done. The data are then sent to SSC for science data processing, science product generation and archiving, as shown in Figure 4.

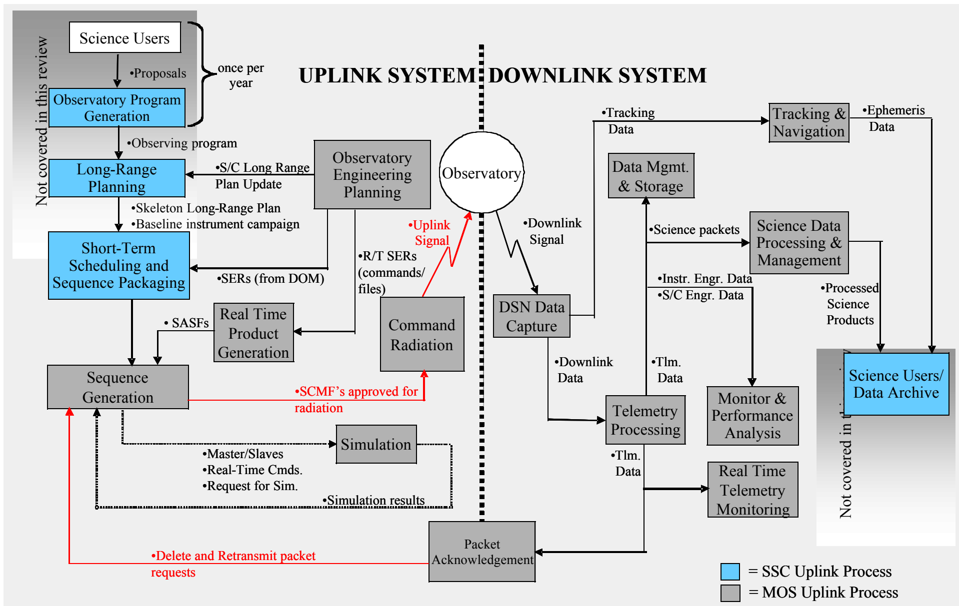

Figure 2. *Spitzer* data flow (Scott et al. 2006).

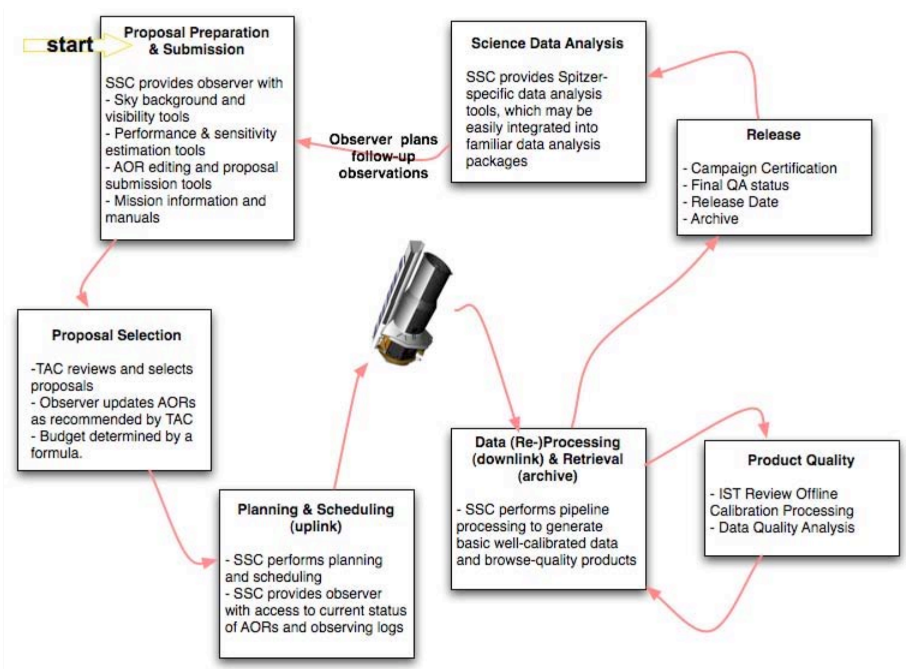

Figure 3. SSC Operations Flow

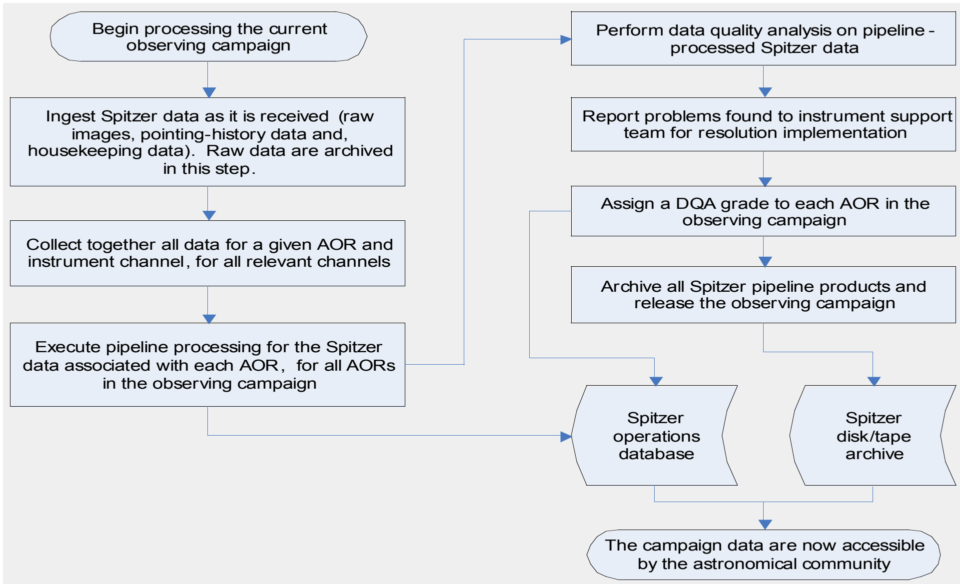

Figure 4. Data (re-)processing and DQA Activities (Mannings et al. 2006)

Spitzer is designed to operate autonomously for 12 hours at a time – this is called a PAO, or Period of Autonomous Operations. An observing sequence is formally referred to as an Astronomical Observation Request (AOR) in *Spitzer* parlance. An AOR is composed of multiple Data Collection Events (DCEs), where a DCE constitutes one or more raw science images obtained with a *Spitzer* instrument (depending on the instrument and observation mode). Right after an AOR finishes, an observer has to wait at least until the end of the PAO before the data come down to the ground.

Assuming all the data come through without errors in transmission (which is usually the case), it takes several hours for data from a PAO to move from the DSN through JPL to the SSC. If there are errors, there are two more opportunities for the data to be transmitted, during the next two downlinks. If the data has not been received by the third downlink, it will most likely not ever come down. There is reprocessing of the observation upon each receipt of additional data.

Since some instrument calibration tasks occur both at the start and at the end of a campaign, before the data are ready for the observer to work with, an observation must make it through end-of-campaign reprocessing. If the observation occurs at the beginning of a 14-day MIPS or IRS campaign, for example, then the observer will have a minimum 14 day wait before this end-of-campaign (EoC) reprocessing begins, and it may take several days to finish.

The DCE FITS images, pointing-history files and housekeeping-data files are ingested at the SSC as they are received from JPL. This involves three basic steps: 1) registering the files in the *Spitzer* operations database, 2) storing the files in the *Spitzer* archive, and 3) applying read-only file permission. The DCEs are then automatically processed and calibrated using a suite of software pipelines that were developed and are periodically upgraded by the SSC downlink software group. Since the housekeeping data are not needed for the routine processing, they are archived in a file system.

The data calibration has two phases. New data are initially processed using calibration data know at the time of processing. This occurs shortly after ingesting and collecting together the data to form a complete AOR. Once all AORs in the entire instrument campaign have been initially processed, the campaign's full set of calibration data is then used to fine-tune the calibration of all the data acquired in that campaign. This is accomplished by a second round of processing, typically within a week of the end of an observing campaign.

The pipeline processing is done on copies of the DCEs (the original DCEs received from JPL are kept pristine in the *Spitzer* archive). The SSC downlink software system generates via pipeline processing for each DCE a Basic-Calibrated-Data (BCD) product, which includes image data that have been converted to absolute flux-density units (MJy/sr or Mega-Janskys per steradian) by the flux-calibration process, and WCS-projection parameters written to their FITS headers as derived from the pointing-history data.

After BCD processing, ensembles of BCDs are further pipeline processed en masse to create so-called post-BCD data products. Typically an ensemble is comprised of all DCEs of the same instrument channel in an AOR, where an instrument channel corresponds to image data limited to a specific infrared spectral passband. However, in some cases, ensembles span multiple instrument channels, such as when computing a common World Coordinate System (WCS) reference frame for re-sampling BCD images, which is necessary for multi-color combining. Examples of post-BCD data products are as follows: re-sampled, co-added images to reduce image-data noise; mosaics of re-sampled, co-added images for mapping areas of the celestial sky that are much larger than an instrument's field of view; catalogs of point sources derived for separate instrument channels; and band-merged lists of point sources detected in common across multiple instrument channels. These advanced data products, as generated directly by SSC automated pipelines, are of sufficient quality to be published directly in peer-reviewed astronomy and astrophysics journals.

Throughout the ingest stage, the BCD-pipeline-processing stage, and the post-BCD-pipeline processing stage, several quantities that are useful in the data quality analysis process are computed and/or stored in various tables in the *Spitzer* operations database. For example, information about missing lines of image data and missing ancillary-data packets (ancillary data are packaged in the DCE's FITS header) is stored in a database table when the DCE is ingested. At the beginning of the BCD pipeline, various image statistics for the DCE are computed and associated with the best instance of the DCE via a database index when stored in the database. At the end of the BCD pipeline, similar image statistics for the associated BCD are computed and associated with a data-product index when stored in the database. Finally, at the end of the post-BCD pipeline, image statistics, outlier counts, and other DQA measures are computed for the post-BCD products and associated with an ensemble-product table when stored in the database. A general image software program called QATOOL is used in the pipelines to routinely compute image statistics.

The five subsystems/segments of the software system and their functions are illustrated in Figure 5. The core functions of each subsystem are:

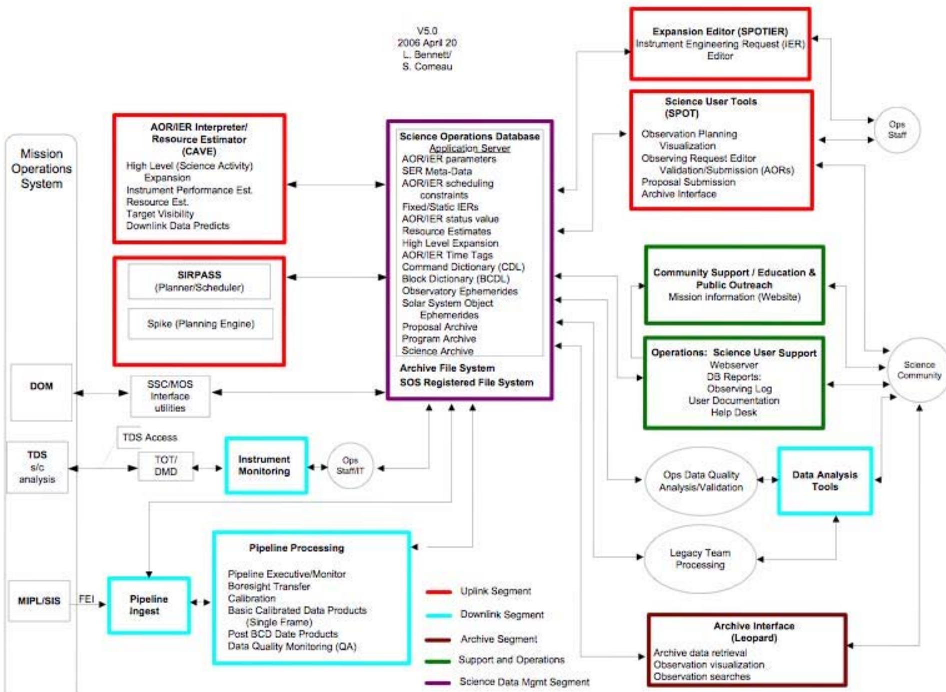

Figure 5. SSC SOS at launch (Bennett et al. 2006).

- Downlink: data ingestion, pipeline processing, data quality analysis.
- Uplink: planning/scheduling, resource planning, science user tools.
- Archive: Leopard.
- Science Data Management: SSC dbms (sodb), public dbms (archive - replicated sodb), proposal dbms (staging, proposals), file management system.
- Support and Operations: science user support, community support, educational outreach, public affairs.

Figure 6 shows the configuration of the SSC hrdware at launch. Being able to read Figure 6 without straining is not the point here—the important thing to note is that the computers and networks were in place for launch. As a risk reduction choice, all the launch computer systems were provided by Sun Micro Systems and all the network equipment was provided by Cisco. This allowed SSC to go to *o*ne vendor in order to address systems or networking troubleshooting and maintenance. At launch, there were vendor spares stored on site; a Sun field maintenance person was on site; and maintenance for both Sun and Cisco had been elevated formally to Platinum, plus SSC had been placed on special status for the days around launch.

Of other interest may be the hierarchical storage management system. Limited pre-launch funding restricted SSC's ability to purchase on-line storage. The engineering solution was to field an HSM with 25 TB capacity - more about this later.

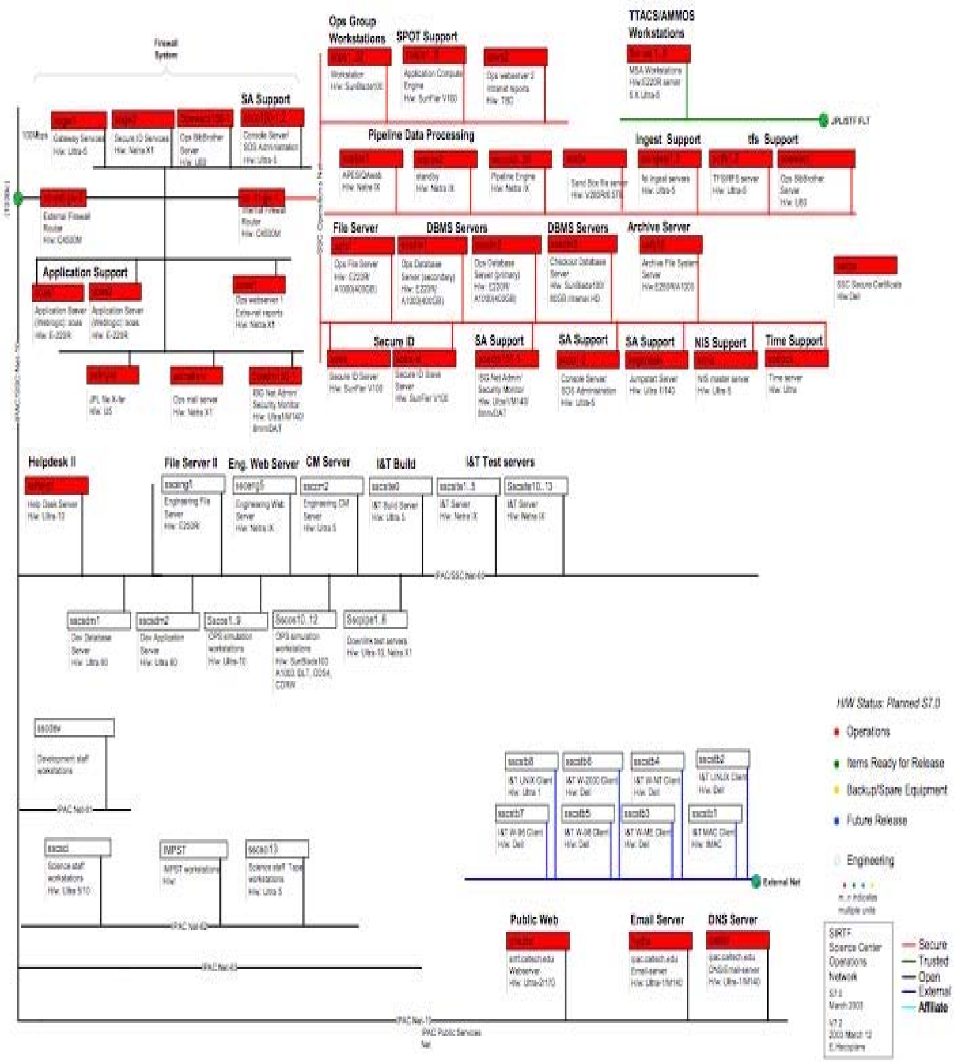

Figure 6. SSC launch hardware.

4. Changes in the First 1100 Days

All the changes were driven by an improving understanding of the SSC Science Operations Systems. Given the system architecture as described previously, these changes could be effected as mostly independent improvements. The changes included migrating from direct-attached storage to network-attached storage with 2Gb fiber channel connections; replacing all pipeline nodes with multi-cpu systems; upgrading the SSC operations network from 100Mb/s to 1Gb/s; changing the network topology from numerous local switches to direct run connecting to a new core switch; upgrading the DBMS system from Informix 9.4 to Informix 10.0 to enhance replication; off-loading obsolete dbms and file system content; and the post-launch development of the public archive interface (Leopard).

5. What are Today's Issues?

With the end of the cryogenic mission two years away, there are still both technical and mission challenges to be met. The most important include the extended mission and its additional technology refresh; should there be a change to an Oracle dbms from Informix; conversion from SAM-FS/QFS to the Solaris 10 ZFS file system; explosive archive growth; file system management where the number of files is three to five billion; managing the currency of data products; dbms (sodb/archive) and file systems; off-load of obsolete data from the operational DBMSs and file systems; upgrade of storage network from 2 Gb to 4 Gb; support for 10 Gb network throughput; and reprocessing and publication of final standard products to the community via IPAC's Infrared Science Archive (IRSA).

6. What Did We Get Right? Get Wrong?

6.1. What Did We Get Right?

The system remained operational in spite of growth of the data volume by more than two orders of magnitude from the original estimates. Hardware and software partitioning and scalability enabled focused improvements, changes and upgrades. The as-built pipelines proved to be efficient. Representative processing times from end-of-campaign to release of data including calibration file selection, reprocessing and archiving were 8 to 10 days for IRAC; 9 to 11 days for MIPS, 16 days for IRS.

6.2. What Did We Get Wrong?

The definition and forecast of the data products was inaccurate. There was unplanned dbms and file system growth since SSC explicitly chose to record all relevant information. There were limited IT resources in the areas of environment and staffing. Risk-based (re-) testing was used.

7. What Would We Do Differently Next Time?

If we changed in only four areas, the improvements would be significant. These areas are:

1. Increase the formality of the engineering lifecycle and its processes.
2. Improve accuracy and fidelity for intermediate and final data products.
   - Improve knowledge of as-built through better configuration management.
   - An all-disk solution for on-line storage, relegating SAM-FS/QFS to archive only.
3. Service level agreements with *Spitzer* project and instrument teams.
4. Improve network architecture
   - Perimeter network for DMZ services
   - Improve internal network providing services to the operations dark network

Acknowledgments. This work was performed at the Infrared Processing and Analysis Center at the California Institute of Technology under contract to the National Aeronautics and Space Administration.

As there is insufficient space to list all the coauthors on the title page, here is the complete list: T. Nguyen, S. Lo, R. Ebert, L. Amy, J. Lampley, J. Jacobson, S. Schurr, C. Smay, E., J. Kakumasu, A. Hui, J. Llamas, J. White, T. Wang, D. Wittman, M. Wittman, R. Scholey, L. Fung, and J. Higgins.

References

Bennett, L., Comeau, S., Levine, D. A. & Dodd, S. R. 2006, Proc. SPIE, 6270, 37B

Kaase Initiatives LLC 2005, Briefing: *SE Tutorial on System Architectures*

Mannings, V., & Laher, R. R. 2006, Data Quality Analysis at the Spitzer Science Center, Proc. SPIE, 6270, 64M

Roellig, T. L., et al. 2004, Proc. SPIE, 5487, 18H

Scott, C. P., Kahr, B. E., & Sarrel, M. A. 2006, SPIE, 6270, 1B

Storrie_Lombardi, L. 2005, Briefing: *Spitzer Science Center and Operations*

Werner, M. W., et al. 2004, ApJS, 154, 25

Werner, M. W. 2006, Proc. SPIE, 6265, 7